10. Sampling and Estimation

Sampling

Sampling may be defined as the selection of same part of an aggregate of totally on the basis of which a judgment about aggregate. Sampling is a process of selecting samples from a group or population to become the foundation on estimating and predicting outcome of the population. Sample is part of population.

Sample survey means study of unknown population on the basis of representative sample drawn from it.

Some Important Terms

1. Population or Universe : *From the statistical Point of view the term "Universe" refers to the total of the items or units in any field of inquiry where the term "population" refers to the total of items about which information desired.* We do not find any difference between population and universe and as such two terms are taken as interchangeable.

The aggregate of such units is generally described populations. The population or universe can be finite and infinite.

2. Elementary Units or Unit : *A well-defined and identifiable object of a group of objects in the populations with measurement or counts are associated is called elementary units.*

3. Observation : *Observation, in which the scientist observes what is happening, collects information, and studies facts relevant to the problem.* In this stage, statistics suggests what can most advantageously be observed and how data might be collected.

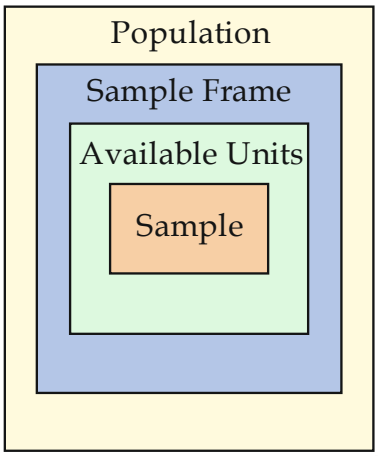
4. Observational Unit : *The unit about which the information has to be obtained is called observational unit.*

5. Sample : *It is part of a population and it is selected with a view to represent. A sample is a finite part of a statistical population whose properties are studied to gain information about the whole.* When dealing with people, it can be defined as a set of respondents (people) selected from a larger population for the purpose of a survey. It is part of population

6. Sample Units : *A decision has to be taken concerning a sample unit before selecting sample.* Sampling unit may be a geographical one such a state, district, village etc.

7. Sample Frame : Sample Frame is a detailed and complete list of sampling units. The list of people from whom draw sample, such as a phone book or 'people shops in town today', may well be less than the entire population and is called a sample frame. This must be representative of the population otherwise bias will be introduced.

Sample frames are usually much larger than the sample. They are used because of convenience and the difficulty of accessing people outside this frame (for example those without a telephone).



8. Statistics(s) and Parameter : *A statistics is a characteristic of sample whereas a parameter is a characteristics of a population.* Thus, when we work out certain measure such as mean, median, mode or the like ones from sample. They are called statistics. For they describe characteristics of a sample.

But when such describe the characteristics of a population they are known as parameters. *We denote Population means (μ) and sample means (x) is a statistics.*

Measure	Parameters	Statistics
Size	N	n
Mean (x)	μ	$\bar{x}$
Standard Deviation	σ	S
Proportion	P	p

9. Sampling Error : The part of the difference between a population value (Population Parameter) and estimate value there of derived from a random sample (Statistical Property of Sample called Sample Statistic). *A sampling error is limited to any differences between sample values and universe values that arise because the entire universe was not sampled.*

Sampling Error (S.E.) = Statistics

Thus these error would not be present in a complete enumeration survey. However these error can be controlled. The modern theory of sampling help designing the survey in such manner that sampling error can be made small and limited.

Some of the Causes for Error in Sampling

- Error in selection of the sample
- Bias in reporting of data
- Diversity of Population
- Substitute of sampling units for convenience
- Faulty demarcation of sampling universe

Types of Sampling Error

(i) Biased Error : *The Errors arise from any bias in selection and estimation etc. these errors creep in because of the bias of the persons or instruments involved in the collection of data, and the nature of statistical method followed is such that the error has a tendency to grow in magnitude with the increase in the number of observation.*

These error arises on account respondent's bias, bias due to non- response, bias in the technique of approximation etc.

(ii) Unbiased Error : *These errors arise due to chance difference between the members of population included in the sample and those not included.* An error in statistics is the difference between the value of a statistics and that of the corresponding parameter.

The main purpose of statistical methods is to avoid the biased error and devise methods in such a manner that the errors if it is only the unbiased ones. The magnitude of such errors is automatically reduced by enlarging the size of observation.

10. Non Sampling Error : *The Non-Sampling Error is the statistical error that arises due to the factors other than the ones that occur when the inference is drawn from the sample. Error due to recording observation basis on the part of the enumerators wrong faulty interpretation of data etc.*

Non-Sampling error can occur at every stage of planning and execution of the census survey. Such errors can arise due to number of causes such as defective method of data collection and tabulation faulty definition, inadequate and inconsistent objectively error in location of units, lack of trained and qualified staff of investigators and errors in response etc.

The non-sampling errors are unavoidable in census and surveys. The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both - sampling errors as well as non-sampling errors. Non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample.

Non-Sampling Error may Arise Due To

- Faulty Sample Plan
- Lack of trained and qualified investigators
- In accuracy in response collected due to bias of respondent or the researcher.
- Errors in design of the survey
- Errors in compilation or publication.

Difference between Sampling and Non-Sampling Error

- *Sampling Error arises on account of fluctuations in sampling whereas Non-Sampling error arises due to other reasons.*
- *Sampling error are not present in census investigation whereas Non-sampling errors affect both type of investigation- census as well as sample.*
- *Sampling error are compensated by enhancing the size of the sample but Non-Sampling errors are of cumulative nature. As such Non-Sampling errors will increase with the increase in size of sample.*

$$\text{Sampling Error} = \pm\sqrt{\frac{2500}{\text{Sample Size}}} \times 1.96$$

Example : What is the sampling error if average weight of the 60 men is 58 kg ?

Solution : Given : sample size = 60

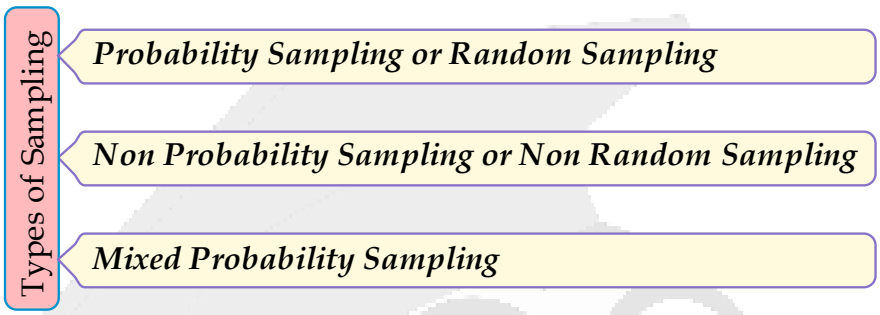
The sampling error is given by

Sampling error can be found out using the formula :

$$\text{Sampling Error} = \pm\sqrt{\frac{2500}{\text{Sample Size}}} \times 1.96$$

Sampling error can be found out using the formula:

$= \sqrt{} \pm 2500/60 \times 1.96 = \pm 12.65$



1. Probability Sampling or Random Sampling

Each person in the universe has an equal probability of being chosen for the sample and every collection of persons of the same has an equal probability of becoming the actual sample.

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population. If for some reasons, the sample does not represent the population, the variation is called a sampling error.

The units are selected independent of each other i.e. when each unit of population has equal chance of selection. It can be simple random sampling with replacement (WR) or Without Replacement (WOR).

(a) Sampling with Replacement : If the unit are drawn one by one each unit after selection is returned to the population before next unit is being drawn so that composition of the original population remain unchanged at any stage of the sampling therefore the sampling procedure is known as simple random sampling without replacement (WR)

(b) Sampling Without Replacement : If however once the units selected from the Population one by one are never returned to the population before the next drawings made then the sampling is known as sampling without replacement.

It is used when

- Population is very large
- Sample size is not very small
- Population is not heterogeneous i.e. Not match variability among the members of the population.

(c) Lottery Method : This is most popular method and simplest method. In this method all the items of the universe are numbered on separate slips of paper of same size, shape and color. They are folded and mixed up in a drum or a box or a container. A blindfold selection

is made. Required numbers of slips are selected for the desired sample size. The selection of items thus depends on chance.

(d) Random Number Table Method : As the lottery method cannot be used when the population is infinite, the alternative method is using of table of random numbers. There are several standard tables of random numbers. But the credit for this technique goes to :

- Prof. LHC. Tippet (1927). The random number table consists of 10,400 four-figured numbers.
- They are fishers and Yates (19380 comprising of 15,000 digits arranged in twos.
- Kendall and B.B Smith (1939) consisting of 1, 00,000 digits grouped in 25,000 sets of 4 digit random numbers, Rand corporation (1955) consisting of 2, 00,000 random numbers of 5 digits each etc.,
- Computer generated Random Sampling Number

(e) Stratified Sampling : *Stratified sampling is a type of sampling method in which the total population is divided into smaller groups or strata to complete the sampling process.* The strata is formed based on some common characteristics in the population data. After dividing the population into strata, the researcher randomly selects the sample proportionally.

Stratified sampling is a common sampling technique used by researchers when trying to draw conclusions from different sub-groups or strata. The strata or sub-groups should be different and the data should not overlap. While using stratified sampling, the researcher should use simple probability sampling. The population is divided into various subgroups such as age, gender, nationality, job profile, educational level etc. Stratified sampling is used when the researcher wants to understand the existing relationship between two groups.

Stratified Sampling is not Advisable

- The population is not large
- Some prior information is not available.
- There is not much heterogeneity among units of populations.

There are two type of allocation of sample size when there is prior information there is no much variation between the strata variance. We consider "proportional allocation or bowley's" allocation where the sample size for different strata are taken as proportional to the population size.

When Strata variance differ significantly among themselves we have taken resources to "Neyman's" allocation "where sample size vary jointly with population size.

(f) Cluster or Block Sampling : It involves arranging elementary items in a heterogeneous population into homogenous subgroup that are representative of the overall populations.

A simple random sample's in which each sampling unit is a collection or cluster, or elements. For example, an investigator wishing to study students might first sample groups or clusters of students such as classes or dormitories, and then select the final sample of students from among clusters. It is also called area sampling.

(g) Multistage Sampling : Multistage sampling divides large populations into stages to make the sampling process more practical. A combination of stratified sampling or cluster sampling and simple random sampling is usually used.

In simple terms, in multi-stage sampling large clusters of population are divided into smaller clusters in several stages in order to make primary data collection more manageable. It has to be acknowledged that multi-stage sampling is not as effective as true random sampling; however, it addresses certain disadvantages associated with true random sampling such as being overly expensive and time-consuming.

- Choosing sampling frame, numbering each group with a unique number and selecting a small sample of relevant discrete groups.
- Choosing a sampling frame of relevant discrete sub-groups. This should be done from relevant discrete groups selected in the previous stage.
- Repeat the second stage above, if necessary.
- Choosing the members of the sample group from the sub-groups using some variation of Probability sampling.

Ordinary multistage Sampling is applied in big inquiries extending to a considerable large geographical area etc.

2. Non Probability Sampling or Non Random Sampling

Non-probability or Non Random sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected.

(a) Judgment Sampling : In this type of sampling items for the sample are selected deliberately by the researcher. Sample drawn entirely on the personal judgment of the investigation. Judgmental sampling is more commonly known as purposive sampling. In this type of sampling, subjects are chosen to be part of the sample with a specific purpose in mind. It is purely subjective and varies from person to person.

(b) Convenience Sampling : In this type of sampling, researchers prefer participants as per their own convenience. The researcher selects the closest live persons as respondents. In convenience sampling, subjects who are readily accessible or available to the researcher are selected.

(c) Quota Sampling : Quota sampling is a sampling methodology wherein data is collected from a homogeneous group. It involves a two-step process where two variables can be used to filter information from the population. It can easily be administered and helps in quick comparison.

Quota sampling is a Non Probability sampling or Non Random Sampling which size of the quota for each stratum is generally proportionate to size of that stratum in population.

Quota sampling is used when the company is short of time or the budget of the person who is researching on the topic is limited. Quota sampling can also be used at times when detailed accuracy is not important. To create a quota sample, knowledge about the population and the objective should be well understood so that the researcher can choose the relevant stratification; next is to calculate quota from each section of the population and at the end keep on adding samples until the quota for each section is met.

(d) Sequential Sampling : In this type of sample is not fixed in advance but it is decided as the sampling process takes place depending on the results of the first sample. A number of sample lots are drawn in sequence one after another from the population depending on the results of the earlier sample.

(e) Snowball Sampling : The Snowball Sampling is a non-random sampling technique wherein the initial informants are approached who through their social network nominate or refer the participants that meet the eligibility criteria of the research under study. Thus, this method is also called as **the referral sampling method or chain sampling method.**

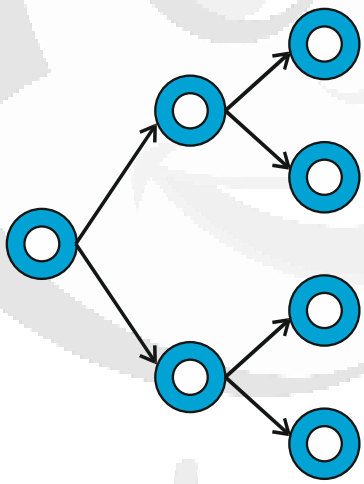
The snowball sampling method is extensively used in the situations when the population is unknown and rare, and it is hard to select the subjects therefrom.

For example, the group of people suffering from Cancer is limited and often reluctant to disclose their disease. And in such case, if the interviewer wants to know how the life of these people have changed due to Cancer, might approach those acquaintances who can refer those individuals who can potentially contribute to the study.

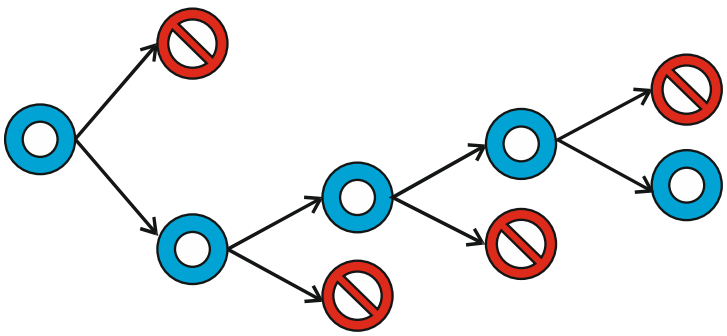
There are following three patterns of snowball sampling :

1. **Linear Snowball Sampling :** Formation of a sample group starts with only one subject and the subject provides only one referral. The referral is recruited into the sample group and he/she also provides only one new referral. This pattern is continued until the sample group is fully formed.

2. **Exponential Non-Discriminative Snowball Sampling :** The first subject recruited to the sample group provides multiple referrals. Each new referral is explored until primary data from sufficient amount of samples are collected.



3. **Exponential Discriminative Snowball Sampling :** Subjects give multiple referrals, however, only one new subject is recruited among them. The choice of a new subject is guided by the aim and objectives of the study



3. Mixed Probability Sampling

Systematic Sampling

It is type of mixed sampling because systematic sampling is partly probability sampling in the sense that the first unit of sample selected probabilistically and partly Non-Probability in the sense that the remaining units of the sample are selected according to a fixed rule which is Non-Probabilistic in nature.

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point and a fixed periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size. Despite the sample population being selected in advance, systematic sampling is still thought of as being random if the periodic interval is determined beforehand and the starting point is random.

Methods of Sampling

If we use randomly sampling method then large sample should be used but stratified method use we select small sample size.

- The problem of response
- Response of problem is not sufficient then we select large sample.
- Availability of response
- Sufficient Resources- Large Sample
- Insufficient Resources - Small Sample

Degree of Freedom

The number of independent observations which make up a statistic is known as the degrees of freedom (d.f.) associated with that statistic. Degree of freedom is the number of values in the final calculation of a statistic that are free to vary. In general, d.f. of a statistic = Number of independent observation of a statistic that are free to vary. In general, d.f. of a statistic = Number of independent observations – Number of parameters estimated. For example, d.f. of the statistic $\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$ is (n – 1), while d.f. of the statistic $\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$ is n.

Standard Error

We have seen that different samples of the same size from the same population will yield different values of statistic under consideration, say sample means. A measure of the variability in different values of sample mean is given by the Standard Error of the sample mean. Standard error of a statistic is the standard deviation of its sampling distribution. Standard error plays an important role in statistical hypothesis testing and interval estimation. Standard error gives an idea about the reliability and precision of the estimate. Smaller the standard error, greater the uniformity of sampling distribution and hence, greater the reliability of the estimate. Also, standard error decreases when sample size is increased. Standard errors of some important statistics are given below:

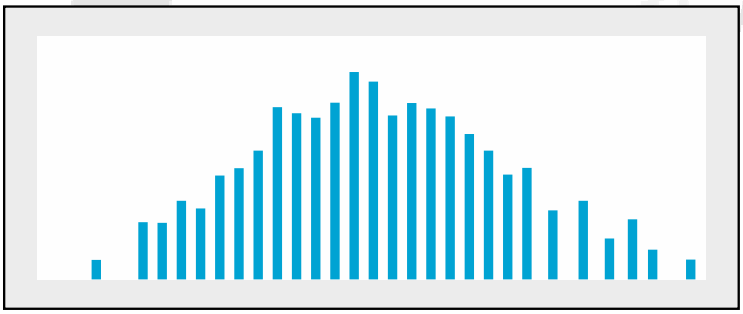
Statistic	Standard Error
Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1$	$\sigma/\sqrt{n}$ (σ is population standard deviation)
Sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma^2\sqrt{2/n}$
Sample variance $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma^2\sqrt{2/(n-1)}$
Sample proportion $p = x/n$	$\sqrt{\pi(1-\pi)/n}$

As can be noticed that standard error depends on (usually) unknown parameters. In order to get some approximated value of standard error, unknown population parameter can be replaced by some feasible estimates.

Sampling Distribution

Distribution of a statistic may not be the same as the distribution of population. We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure.

Here, it is important to discuss normal distribution briefly. Normal distribution is represented by the bell shape of the histogram of given data-set as below :



In many statistical analysis, it is important that the given data has normal distribution. Normal distribution is denoted by N(μ, σ), where μ is mean and σ is standard deviation of the distribution. Normal distribution can be used to model any variable taking values – ∞ to ∞, provided the values on that variable have a bell shaped histogram. Reader may refer to some standard text on statistics to know more about normal distribution.

Some commonly used sampling distributions are given below:

Sampling Distribution of Mean

Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, N (μ, σ), the sampling distribution of mean would also be normal

with mean $\mu_{\bar{x}} = \mu$ and standard deviation $= \sigma\sqrt{n}$, where μ is the mean of the population, σ is the standard deviation of the population and n means the number of items in a sample, but when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., N (0, 1), we can write the normal variate $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ for the sampling distribution of mean.

This characteristic of the sampling distribution of mean is very useful in several decision situations.

Sampling Distribution of Proportion

Proportion is a measure of attribute. Let us consider that the population is divided into two mutually exclusive and collectively exhaustive classes—one class possessing a particular attribute while other class not possessing that attribute. For example, people in a city could be divided into "Smokers" and "Non-smokers". Let N = population size, X = number of people out of N possessing a particular attribute, π = X/N = actual proportion of the people possessing the specified attribute. Let a sample is selected from this population with n = sample size, x = number of people in the sample possessing the specified particular attribute, p = x/n = sample proportion. Note that, X and P are population parameters, while x and p are sample statistics. Also, p provides an estimate of π. It can be showed that the distribution of x is Binomial (n, π). Using the property of binomial distribution, for sufficiently large sample size n, we have

$$Z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$$

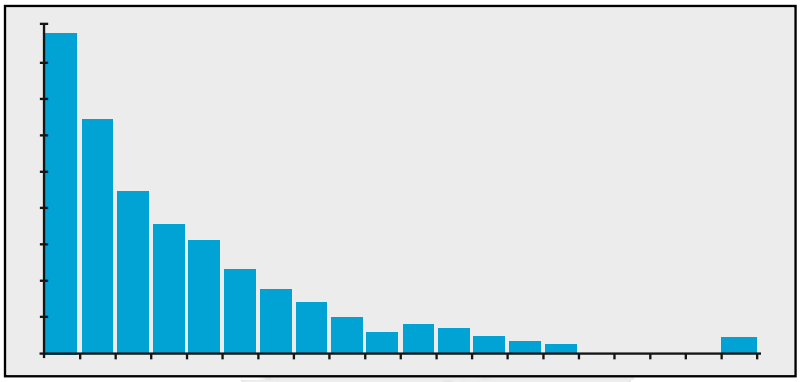
Practically, this result is true for $n \geq 30$ or, when $nx \geq 5$ as well as $n(1 - \pi) \geq 5$.

Central Limit Theorem

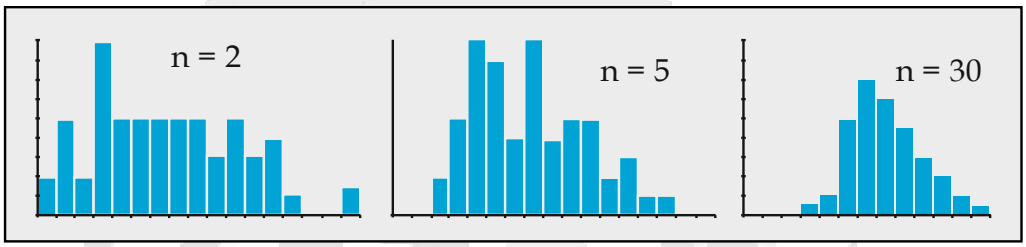
When sampling is from a normal population, the means of samples drawn from such a population are themselves normally distributed. But when sampling is not from a normal population, the size of the sample plays a critical role. When n is small, the shape of the distribution will depend largely on the shape of the parent population, but as n gets large (n > 30), the shape of the sampling distribution will become more and more like a normal distribution, irrespective of the shape of the parent population. The theorem which explains this sort of relationship between the shape of the population distribution and the sampling distribution of the mean is known as the central limit theorem. This theorem is by far the most important theorem in statistical inference. It assures that the sampling distribution of the mean approaches normal distribution as the sample size increases. In formal terms, we may say that the central limit theorem states that the distribution of means of random samples taken from a population having mean g and finite variance σ^2 approaches the normal distribution with mean μ and variance σ^2/n as n goes to infinity.

"The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population.

For example, 1,800 randomly select values from an exponential distribution are plotted below :



Various samples of sizes 2, 5, and 30 are drawn this population sample mean values are obtained. The distributions of the sample means for various sample sizes can be observed from following plots :



It can be noted that for sample size 30 the shape of sampling distribution is almost normal.

Point Estimation

Point estimate is a single valued estimate of an unknown parameter. The statistic $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_2$ (sample mean) is used to estimate population mean μ. So, sample mean is a point estimate of the population mean, The statistics $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ are used to estimate known population variance σ^2. Therefore these two statistics s^2 and s_1^2 are point estimates of σ^2. Conventionally, $\bar{x}$, s^2 and s_1^2 are called as estimators and specific values (such as $\bar{x}$ = 20, s^2 = 4) are called as estimates of the parameters. An unknown parameter may be estimated by more than one estimator. For example, σ^2 is estimated by s^2 and s_1^2.

Ques. *Match the items of List-I and List-II and indicate the code of correct matching of the items :*

	List-I		***List-II***
(a)	$\frac{z^2.\sigma_p^2}{e^2}$	*(i)*	*Measurement for Kurtosis*
(b)	$\frac{\lvert \bar{X}_1 - \bar{X}_2 \rvert}{\sqrt{\sigma_p^2 \frac{1}{n_1} + \frac{1}{n_2}}}$	*(ii)*	*Calculated value of F ratio*
(c)	$\frac{\mu_4}{\mu_2^2}$	*(iii)*	*Statistical approach to find out the size of sample*
(d)	$\frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$	*(iv)*	*Calculated z value of mean differences*

Codes : ***(NTA UGC-NET Dec. 2015 P-III)***

	(a)	***(b)***	***(c)***	***(d)***
(1)	*(i)*	*(ii)*	*(iii)*	*(iv)*
(2)	*(ii)*	*(iv)*	*(iii)*	*(i)*
(3)	*(iii)*	*(iv)*	*(ii)*	*(i)*
(4)	*(iii)*	*(iv)*	*(i)*	*(ii)*

Ans. *(4) Correct match is given below :*

	List-I		***List-II***
(a)	$\frac{z^2.\sigma_p^2}{e^2}$	*(i)*	*Statistical approach to find out the size of sample*
(b)	$\frac{\lvert \bar{X}_1 - \bar{X}_2 \rvert}{\sqrt{\sigma_p^2 \frac{1}{n_1} + \frac{1}{n_2}}}$	*(ii)*	*Calculated z value of mean differences*
(c)	$\frac{\mu_4}{\mu_2^2}$	*(iii)*	*Measurement for Kurtosis*
(d)	$\frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$	*(iv)*	*Calculated value of F ratio*

Ques. *If the population is heterogeneous, which one of the following probability sampling methods is more appropriate ?* ***(NTA UGC-NET June 2015 P-II)***

(A) Sequential sampling *(B) Quota sampling*

(C) Double sampling *(D) Stratified sampling*

Ans. *(D) If the population is heterogeneous, then Stratified sampling methods is more appropriate.*

Ques. *Which one of the following is* ***not*** *the* ***correct*** *statement regarding sampling distribution of mean?*
(NTA UGC-NET June 2015 P-III)

(A) *Sampling distribution of mean is normally distributed for large sized samples.*

(B) *Sampling distribution of mean is normally distributed for small sized samples drawn from not normally distributed population.*

(C) *'t' distribution is not normally distributed.*

(D) *Mean of the sampling distribution of mean is equal to the parametric value of mean.*

Ans. *(B)* *Statement in option B is incorrect regarding sampling distribution of mean.*